Pipelines Overview

Session 3 Brock Tibert

ExamSoft

- You will need to make sure you don't have issues with ExamSoft/Examplify
- 9/18 due date
- Please use that to ensure that you can successfully install and complete the assessment
- Student ExamSoft guide
- questromhelp@bu.edu for help/assistance

– https://support.examsoft.com/hc/en-us/articles/11146797283469-Examplify-Download-the-Installers-for-Windows-Mac

- Code is under Quizzes and Tests

Groups

● At the very least, start

coordinating in your groups to find common times to meet/discuss status each week

● Also should start thinking

about what you collectively want to build this semester

● As we move into hands-on

demos, this can start to frame out your work

● Will have an exercise in groups

today

Today's Agenda - Hands On Demos and Practice

- Hands-on Demo: ERD Design via LucidChart (can be anything, really)
- Hands-on Recap: GCP Console Refresher

- Console refresher
- Cloud shell/Terminal to simplify CLI and API commands (user-specific)
- Develop against a Github project

- Hands-on Recap/Demo: BigQuery Schema management 101
- In-Class Project: AWS Blogs Design Thought Exercise
- Technical Site overview for labs and code samples to help your projects!
- Quiz will be released tomorrow

- A small low-stakes quiz (MC, matching)
- Review the concepts we have covered
  - Data flows and pipelines
  - Considerations
  - Targeted at reinforcing core concepts and intuition before we start to build out our stack!

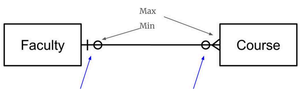
ERDs 101 / Schema Design

ERDs - Crow's Foot Notation

Minimum and Maximum Cardinality

● There is often a "minimum" cardinality and "maximum" cardinality

defined

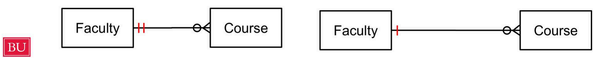


Each course is taught by 0 (at minimum) or 1 (at maximum) faculty

Each faculty teaches 0 (at minimum) or many (at maximum) courses

● To denote "exactly one," you can either have two vertical bars or simplify

it and just have one vertical bar:



Brief Demo ERD for Schema Design

● Concepts > Tools ● Lucidchart -> Free academic

license (if you really need it) ● Mermaid: syntax-based

diagrams; plays well with automation frameworks and

VCS

Fundamentals Hands Development in Cloud Console and BigQuery

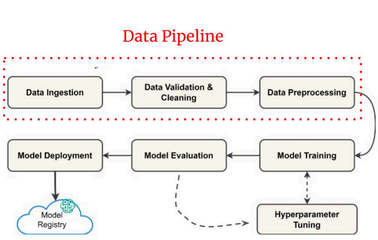
Notes/Terminology

- SQL and terminology is not the same for all databases and providers!
- Data types can be different or behave slightly differently (case sensitivity)

BigQuery Quick Tips

- Your project comes with a BigQuery Database
- Datasets == schemas
- Tables == tables
- You can use the console or a database IDE, but sometimes this is an abstraction over an API (GCP vs Redshift == postgres)
- Cloud Consoles are getting to be feature rich, but be careful on depending
  - n them for all of your work (we want to package our work into projects)
- You are not required to use BigQuery or anything I demo. I am using tools to help show intuition. Concepts > Tools

In-Class Project AWS Blogs/Knowledge Base



How does this fit together with later themes?

The data pipeline should feed the ML process. Feature engineering should be baked into the data pipeline?

Why?
- Ensure everyone uses the same definition
- Consistent reporting

AWS Blogs

A framing for our topics and demos this semester

● Can be framed for all three

phases of the project

● Gives us grounding for working through our discussions this semester ● Relatively intuitive and

constrained corpus of data

● Document on Blackboard for

this class session, let's review some links

Let's review a few blog entries

Group Lab

Recall, ideally our initial schema can help us setup all three themes for this course.

- Pipelines for BI/reporting
- ML model off of the data
- LLM/Gen AI

Work in your team groups, how might you setup landing targets for the AWS blog data? Focus on high-level data flow and schema design)

I am going to use these feeds:

'big-data/feed/' 'compute/feed/', 'database/feed/', 'machine-learning/feed/', 'containers/feed/',

'infrastructure-and-automation/feed/',

'aws/feed/', 'business-intelligence/feed/', 'storage/feed/'

- Slides on Blackboard for us to brainstorm as a group. One slide for each team
- Use any tools you want.
- I will call on groups to present

Technical Resources Site

● Used for tutorials and technical

resources where it's easier to provide code samples/context over Blackboard

● As we now are into the build

phase, used for class session labs

● Site will build incrementally

over the semester

● Intended to be dedicated

out-of-class resource for you and your team projects!

● https://tinyurl.com/BA882-fall2024-techsite

Stepping Back: Pipeline Considerations

More advanced needs might require all sorts of different types of tables/models.

Here we are seeing a flow of the data through tables as a segue into next week.

This extent isn't always necessary, but, worth knowing that there are patterns to leverage as projects get more complex!

Start small! → Don't over-engineer out of the gate